# Determining physical constraints in transcriptional initiation complexes using DNA sequence analysis

Ryan K. Shultzaberger [*], Derek Y. Chiang [*], Alan M. Moses [†], and Michael B. Eisen [* ‡§]

version = 1.17 of space.tex 2007 June 29

## 1 Abstract

Eukaryotic gene expression is often under the control of cooperatively acting transcription factors whose binding is limited by structural constraints. By determining these structural constraints, we can understand the "rules" that define functional cooperativity. Conversely, by understanding the rules of binding, we can infer structural characteristics. We have developed an information theory based method for approximating the physical limitations of cooperative interactions by comparing sequence analysis to microarray expression data. When applied to the coordinated binding of the sulfur amino acid regulatory protein Met4 by Cbf1 and Met31, we were able to create a combinatorial model that can correctly identify Met4 regulated genes.

[*]Department of Molecular and Cell Biology, University of California, Berkeley
[†]Graduate Group in Biophysics, University of California, Berkeley
[‡]Department of Genome Sciences, Genomic Division, Ernest Orlando Lawrence Berkeley National Lab
[§]Corresponding author: 1 Cyclotron Road, Mailstop 84-171, Berkeley, CA 94720. mbeisen@lbl.gov

# 2 Introduction

The regulation of transcriptional initiation from individual eukaryotic promoters is often controlled by multiple cooperatively interacting transcription factors. These factors bind to separate sites in cis-regulatory sequences and physically interact with each other, either directly or through additional proteins, to activate or repress transcription [1, 2, 3]. These physical interactions among transcription factors must constrain how their binding sites can be positioned relative to each other and to the relevant promoters. Yet, there is often considerable variability in the order, orientation and spacing of binding sites for interacting transcription factors [4, 5, 6]. Understanding how the arrangement of sites is related to the stability of these complexes and their regulatory activity is essential if we are to understand the regulatory content of eukaryotic genomes.

To successfully model the binding of multi-meric complexes to different target sequences, many energetic contributions need to be considered. The affinity of each transcription factor for DNA varies considerably with the precise bound sequence, even among known in vivo targets [7, 8]. The stability of the entire complex is also dependent on how compatible the positioning of the sites are with the protein-protein interactions necessary to form the complex. Poorly positioned sites presumably introduce clashes or strain into either the complex or DNA which will, in turn, reduce the stability of the complex.

Here, we combine DNA sequence analysis and genome-wide expression data to discern the constraints on the arrangement of binding sites for transcription factors involved in regulating the synthesis of sulfur-containing amino acids in the yeast *Saccharomyces cerevisiae*. This work builds on our previous modeling of bipartite prokaryotic ribosome and $\sigma^{70}$ binding sites [9, 10]. In both of these cases, initiation requires the cooperative binding of two independent components separated by a variable spacer, the Shine-Dalgarno and P site for ribosome binding sites, and the $-10$ and $-35$ for $\sigma^{70}$ binding sites [11, 12, 13, 14]. Since there were a large number of characterized sites for these systems, we constructed a robust distribution of the allowable spacings between binding components. Assuming that the spacing that would induce the least amount of strain in the protein or in the bound DNA upon binding would be the most commonly observed, and that the frequency of occurrence of all other spacings would be directly related to the energetic consequence of using that spacing, we could model the energetic contribution of different spacings to the formation of a stable initiation complex.

Cooperatively acting transcription factors in eukaryotes are similar to the prokaryotic ribosome and $\sigma^{70}$ in that they have independent binding components separated by variable spacers, but they are different in that the components are not physically linked upon binding and therefore can bind in different orders, orientations, and with greater variability in their spacing. We have devised a method to determine these additional physical constraints by optimizing an information theory based model against microarray data. We can use these optimized constraints to not only infer structural characteristics of the regulatory complex, but also to quantify the binding of these multi-meric complexes to different DNA sequences, and to accurately predict target genes.

Met4 is the major transcriptional activator of sulfur utilization genes in *Saccharomyces cerevisiae* even though it does not bind directly to DNA [15, 5]. Met4 stabilization is dependent upon

at least two additional proteins. One of these is the centromere-binding factor (Cbf1) [15], whose DNA binding activity is stimulated by association with Met28 [16]. It has been suggested that the Cbf1-Met28-Met4 complex may be sufficient for activation of some genes, but coordination by a second factor is necessary for others [4]. We are interested in describing this coordinated system. The second stabilizing factor that we will study is Met31, a factor unique to sulfur regulation [17].

Neither the distance between Cbf1 and Met31 in functional Met4 stabilizing complexes, nor the distance between Met4 and the initiating polymerase is fixed [5]. We extended the information theory-based method we used to study prokaryotic translational and transcriptional initiation to model Cbf1 and Met31 interactions, allowing for the greater flexibility present in this system.

# 3 Materials and Methods

## 3.1 Cbf1 and Met31 binding models

We built a weight matrix describing the sequence preferences of Cbf1 from 16 Cbf1 binding sites characterized by Wieland *et al.*[18]. Binding matrices were built using the standard **Delila** programs [19, 20]. Since Cbf1 binds as a homodimer, we used each sequence and its complement to build our model [21] (Fig. 1A). Because of the lack of experimentally verified binding sites for Met31, we modeled its binding by analyzing 21 non-divergently transcribed genes identified in a Met4 chromatin immuno-precipitation assay [3] (we selected genes with $p < 0.001$). We used MEME [22] with the -tcm model and required at least 10 copies of a motif to identify sequences enriched in these target genes, from which we computed an initial Met31 weight matrix. We then scanned the entire genome for sites with greater than 10 bits of information against this model, identifying 209 sites, from which we constructed the Met31 weight matrix used in our analysis (Fig. 1B).

## 3.2 Searching algorithm

Multi-component binding systems with variable spacing between components have previously been modeled [9, 10]. In the case of the prokaryotic ribosome and $\sigma^{70}$, the binding components are physically connected. In both instances, deviations in the optimal spacing between components introduces strain in the bound complex and affects the binding energy [11, 12, 13, 14]. To model these multi-meric binders the following equation was used:

$$\text{Flexible Site Information} = R_i(A) + R_i(B) - GS(d) \quad \text{(bits/site)} \tag{1}$$

where $R_i(A)$ is the relative strength, or individual information, of binding factor A, and $R_i(B)$ is the relative strength of binding factor B according to [20]. $GS(d)$ is the gap surprisal (based on Tribus'

surprisal function [23]), or penalty of having a spacing of $d$ between sites A and B as determined by [9, 10]:

$$GS(d) = -\log_2 \frac{n(d)}{n} + e(n) \quad \text{(bits/spacing)}. \tag{2}$$

$n(d)$ is the number of occurrences at spacing $d$ and $n$ is the number of total occurrences over the allowed values of $d$. $e(n)$ is a small sample correction value [24, 9]. For our initial analysis of Cbf1 and Met31, we used a flat spacing distribution, where all spacings have the same gap surprisal.

For the ribosome and the polymerase, the binding components are physically linked and can only bind in one orientation relative to each other. For cooperatively acting transcription factors though, there could be variation in the orientation of the sites relative to each other. To account for this, we can adapt the gap surprisal function to:

$$OS(o) = -\log_2 \frac{n(o)}{n} + e(n) \quad \text{(bits/spacing)}. \tag{3}$$

where we calculate an orientation surprisal ($OS(o)$) that is the logarithm of the frequency of occurrence at each orientation. For a system where both orientations occur at equal frequency, the number of occurrences at either orientation would be $n(o) = 1$, and the total number of occurrences is $n = 2$. The orientation surprisal for this system would therefore be 1 bit of information. In a system where there is no variability in orientation, the frequency of occurrence at that orientation would be $\frac{n(o)}{n} = 1$, and therefore the orientation surprisal would be 0 bits. The advantage of the $OS(o)$ calculation is that we can model the subtle energetic differences for systems that allow either orientation, but favor one over the other.

To calculate the total information for Met4 coordination, we can now expand equation (1) to:

$$\text{Flexible Site Information} = R_i(Cbf1) + R_i(Met31) - GS_{Cbf1-Met31}(d) - OS_{Met31}(o) \quad \text{(bits/site)}. \tag{4}$$

There is no orientation surprisal for Cbf1. Since Cbf1 is homodimeric and has a symmetric matrix, the Cbf1-DNA complex would be identical for either orientation. In this case, the frequency of occurrence of a given orientation would be 1, and $OS_{Cbf1} = 0$ bits. Therefore, the orientation surprisal only applies to asymmetric binders.

Combinatorial scans were done using **multiscan** [10] to identify and quantify Cbf1/Met31 cooperatively acting binding sites in the genome. The individual information contribution for both sites ($R_i(Cbf1)$ and $R_i(Met31)$) were calculated over the range $-4$ to $+5$, since this is the range of conservation for both logos (Fig. 1) [20]. Sites were only considered if each component had an $R_i > 0$ bits (which would correspond to a $-\Delta G$ of binding [20, 25]) and they have a flexible site information $> 0$ bits. For a site to have a positive flexible site information, the ordering and orientation of the pair have to be within the defined spacing and ordering parameters. For any spacing or orientation outside of the specified range, the sites would have a surprisal penalty equal

to infinity according to equations (2) and (3), and a flexible site information $< 0$ bits according to equation (4).

All genes in the genome were then ranked based on the strength of their strongest upstream site. Microarray expression data for sulfur amino acid pathway-affected cells (see Microarray Datasets) were then averaged for the top 20 genes in our ranking. This was done independently for induction and repression experiments.

The physical constraints that we want to define are: the ordering of the sites relative to the gene start, the orientation of the matrices, the maximum allowed distance between Met4 and the polymerase binding site, and the spacing range between Cbf1 and Met31 that can bind Met4. We varied these constraints, and iteratively refined the model to get the optimal predictor. We evaluated any given set of parameters by calculating the average expression change in the top 20 ranked genes. The greater the expression change the better the model.

Another approach could be to cluster genes based on similar trends in expression data across several experiments, and then try to train our parameters based on this set of genes. One disadvantage of this is that it is difficult to discern directly from indirectly regulated genes in these clusters. By scanning the genome and ranking the genes, we are selecting only for genes that are directly regulated. Also this approach does not exclude genes that are regulated but had anomalous expression data due to experimental error. Since there have been at least 20 genes implicated in sulfur assimilation [5], we choose to average the top 20 gene expression differences to evaluate our model.

### 3.3 Microarray Datasets

We used microarray data from two sources for our analysis. Gasch *et al.* [26] reported amino acid starvation data, where transcription of Met4 regulated genes was induced. Fauchon *et al.* [27] reported $Cd^{2+}$ addition experiments where Met4 regulated genes were induced, and Met4 deletion experiments where Met4 regulated genes were repressed. Our models were optimized against these data as mentioned above. Microarray expression patterns were visualized using **TreeView** [28]. The yeast genome sequence and annotation that we used in our analysis came from Genbank accession numbers NC001133 to NC001148.

## 4 Results

### 4.1 Cbf1 and Met31 logos

Since Cbf1 is a homodimeric protein, we used all sequences and their complements to build our model [21]. Conservation at positions $-2, -1$ and $+2, +3$ is strong and does not match the helical accessibility wave (Fig. 1A). Deviation of sequence conservation from the helical accessibility ⇐Fig 1

wave is generally an indicator of structural changes in the DNA substrate [29]. This may be consistent with the observed bending of DNA by Cbf1 [30].

The Met31 model was built as described in Materials and Methods (Fig. 1B). Sequence conservation appeared to follow the helical accessibility wave well, and it was contained within one major groove. Met31 has an asymmetric binding site, so it can possibly bind with two different orientations. We tested both orientations in our analysis. The information content for the Cbf1 logo is 12.9 bits over the range $-4$ to $+5$. The information content for the Met31 logo is 11.9 bits over the range $-4$ to $+5$.

## 4.2 Orientation and ordering

Since Cbf1 and Met31 are not physically linked upon binding, it was not immediately obvious what the ordering and orientation constraints on their binding are in functional Met4 docking complexes. To determine this, we tested the predictive capabilities of all combinations of orientation and ordering for Cbf1 and Met31 using the gene-ranking approach described in Materials and Methods. Briefly, we determined the flexible information for the cooperative model as determined by equation (4) [9, 10], and ranked all genes in the genome based on the strength of the strongest site in the intergenic region immediately upstream of their starts. We then calculated the average expression fold change of the top 20 genes in this ranking based on Met4 induced and repressed microarray experiments [26, 27]. We regarded those combinations that gave the highest average microarray expression change to be the optimal organization for Met4 coordination. Fig. 2 shows how well different combinations performed. ⇐Fig 2

Cbf1 alone was not sufficient to identify the Met4 regulated genes. The average expression fold change for the top 20 ranked genes was 0.11 and 0.23 for induction and repression data respectively, we report corresponding values for all other combinations. Met31 alone appeared to be a better predictor than Cbf1, but was still weak (0.54 and $-0.85$). By searching for Cbf1 and Met31 sites together, with a maximum spacing of 100 bases between the zero positions of the binding components (Fig. 1) and the downstream component could be a maximum of 1000 bases upstream of the gene start, the prediction was better. If we searched with the order Cbf1-Met31-gene start, we were able to identify more genes with the expected microarray pattern than with the order Met31-Cbf1-gene start, even though the sites appeared to be low in the ranking (0.78 and $-1.49$ vs. 0.28 and $-0.64$).

Since Cbf1 is a homodimer, its binding is independent of orientation. Since Met31 is monomeric, its binding is orientation dependent. When we allowed for both orientations of Met31 downstream of Cbf1, we got the largest change of expression (0.99 and $-1.70$). This suggests that transcriptional activation by Met4 requires a Met31 site with any orientation to fall between Cbf1 and the gene start (bottom right panel of Fig. 2). All models for the remainder of this analysis will have these ordering and orientation requirements imposed on them. The designation of the Met31 model orientation as “normal” or “inverted” is arbitrary. We also tested the “inverted” Met31 model alone, and inverted Met31 upstream of Cbf1, but the results were similar to equivalent scans

with the "normal" orientation (data not shown).

## 4.3 Spacing constraints

There are two spacing constraints on this system, the distance between the Met4 docking complex and the initiating polymerase, and the distance between the two binding components (Cbf1 and Met31) within the Met4 docking complex. To define what these spacing ranges are for functional Met4 binding sites, we systematically modeled different spacing ranges, and quantified the models by the gene-ranking approach previously described. Interestingly, if we varied one of the spacing constraints, the optimal spacing for the other would differ slightly. To identify which spacing parameters define the optimal predictor, we varied both spacings simultaneously, and quantified their predictability by averaging the expression change of their 20 highest ranking genes.

We increased the maximum allowed distance of the Met4 docking complex from the gene start in 50 base increments as measured by the distance between the Met31 site and the translational initiation codon. At each 50 bp increment, we varied the minimum and maximum allowed distance between Cbf1 and Met31 from 1 to 100 bases. These distances are relative to the zero position of both matrices (Fig. 1). We then summed the average expression change for the induction and repression experiments for all combinations of spacings, and determined which combination predicted the microarray data best.

For the first spacing constraint, the distance between Met4 and the polymerase, we found the optimal maximum spacing was 450 bases (Fig. 3). The predictability of the model seemed to increase linearly from 100 to 350 bases suggesting that the sites are evenly distributed over this range. There appeared to be few or no genes with sites closer than 100 bases upstream, or sites farther than 450 bases upstream that had the expected expression pattern. ⇐Fig 3

For the second spacing constraint, the distance between Cbf1 and Met31, we found the optimal spacing range to be $-13$ to $-68$ bases, the minimum to maximum spacing allowed between each site (Fig. 4). This was the range used in the analysis in Fig. 3. Ranges close to $-13$ to $-68$ appeared to have a similar level of predictability as indicated by the redish semi-circle in Fig. 4, but $-13$ to $-68$ had the highest expression change and the tightest range. The average expression changes for these two spacing parameters were 1.39 and $-2.21$ for induction and repression data respectively. ⇐Fig 4

## 4.4 Optimal model

Based on the analysis in Fig. 2, Fig. 3, and Fig. 4, the optimal model is shown in Fig. 5. This model requires a Cbf1 site to be 13 to 68 bases upstream of a Met31 site with either orientation, and for the Met31 site to be no more than 450 bases upstream of the translational initiation codon. When we scanned the genome with this model, we see that most of our top hits are genes known to be involved in sulfur amino acid biosynthesis (Fig. 6). Only two of the genes in the top twenty ⇐Fig 5 ⇐Fig 6

hits, have an unexpected expression pattern (Reb1 and Gar1). Additional analysis of these sites show that they both have a strong Cbf1 site, but a "T" instead of "G" at position $+1$ of their Met31 site. This suggests that the information contribution at position $+1$ may be greater than that in our current matrix. Several genes have both the expected expression profile and a predicted Met4 binding site, but their functions have not been biochemically characterized (DDR48, YIL074C, YJL060W, YHR112C). Clustering of co-regulated genes by the gene-ranking method may have identified other genes involved in sulfur utilization.

To ensure that we did not overfit our model, we used two jackknife tests. In the first, we removed each of the top 20 ranked genes as determined by our optimal model from our gene list, recalculated the optimized parameters without this gene by the gene-rank method, and then scanned and reranked the excluded gene based on the new parameters. We found that the ranking of all genes except Met28 did not change by a ranking greater than 2. Met28 went from a ranking of 5 to 1429, but this was expected because the Cbf1 to Met31 spacing of Met28 was at the maximum spacing allowed of 68 bases in our optimal model, and when removed it fell outside of the newly determined maximum spacing of 64. In the second jackknife, we randomly removed 5 of the top 20 sites from our gene list, determined the optimal spacing parameters, and then reranked the excluded 5 genes based on these new parameters. Over 100 iterations of this, again most of the genes did not show a ranking change of more than 2, but 5 of the genes at extreme spacings showed noticeably larger rank differences.

To test whether there is a tendency for Cbf1 and Met31 to bind on the same face of the DNA, we plotted the relative spacing between the two sites on a cosine wave with the same period as B-form DNA, 10.6 bases (Fig. 5). We plotted the spacing of 18 of the 20 top hits (all sites except for Reb1 and Gar1) and YHR112C and Met10, which had both a strong flexible information and expression change. We determined the phase of the cosine wave that gave the highest average helical location of these 20 spacings, and found the optimal phase to peak at $-13.24$ bases relative to the Met31 zero position. To see if the relative placement of these spacings on the cosine wave is higher than expected, we determined the average helical location of random sets of 20 Cbf1/Met31 pairs. Our set had an average helical positioning greater than 99 percent of random sets.

To calculate the flexible individual information for each binding site, we used equation (4). Since we did not know the energetic effect of different spacings on the complex initially, we treated all spacings equally. That is, over the range 13 to 68 (56 bases of variability) all positions had the same gap surprisal of $GS(d) = -log_2(\frac{1}{56}) = 5.81$ bits according to equation (2). We also assumed an equiprobable occurrence of each orientation of Met31, so that $OS_{Met31}(o) = -\log_2(\frac{1}{2}) = 1$ bit of information according to equation (3). Therefore the *GS* and *OS* variables in equation (4) effectively become constants summing to 6.81 bits of uncertainty for each site. Because of the small number of target genes, and the already strong predictive capabilities of our model, we cannot determine the individual spacing constraints for this system. If we had a system with more sites, robust spacing and orientation distributions could be determined and individual penalties could be assigned.

We can use these values to predict the $R_{sequence}$ or average information content for this system

which is:

$$R_{sequence}(Met4) = R_{sequence}(Cbf1) + R_{sequence}(Met31) - \bar{GS}(d) - \bar{OS}(o) \quad \text{(bits)}. \qquad (5)$$

$\bar{GS}(d)$ is the mean $GS(d)$ value for all sites, and $\bar{OS}(o)$ is the mean $OS(o)$ value. According to this equation $R_{sequence}(Met4) = 12.9 + 11.9 - 5.81 - 1.0 = 18.0$ bits of information.

For each gene we plotted the strength of its strongest upstream Met4 binding site according to the model in Fig. 5 and its average expression change for induction and repression experiments (Fig. 7). At about $R_i > 14$ bits, the number of genes that showed no, or an unexpected expression ⇐Fig 7
difference was significantly lower. This is about the same $R_i$ as the site upstream of Met10 (14.1 bits), the lowest ranking sulfur assimilation protein in our analysis.

# 5 Discussion

Transcriptional initiation in eukaryotes is often regulated by multiple cooperatively acting factors. Often these factors can only positively affect transcription if they physically interact either directly or indirectly through additional proteins with the basal transcriptional machinery. Understanding the physical constraints that determine functional cooperativity is essential for us to be able to model, predict, and engineer genetic control systems. These constraints generally are not rigid, but allow for variability in the arrangement of sites in functional complexes and subsequently there is variability in the stability of the complexes. Here, we have introduced a way to include orientation and order into the information theoretic description of pattern recognition at the promoter. This combined with weight matrix based binding models [20] and spacing constraints [9, 10] gives us quantitative tools to model the sequence basis of eukaryotic transcriptional regulation.

The simplest constraint of Met4 coordination to define is the ordering of the sites within the complex. For Met4, our model matches microarray data poorly when the order is Met31-Cbf1-gene start, but matches well with the order Cbf1-Met31-gene start (Fig. 2). This is consistent with experimentally determined ordering constraints [31]. These results suggest that the Met4-Cbf1 binding surface is distinct from the Met4-Met31 surface, and that the Met4-TFIID binding surface is closer to the Met4-Met31 surface, placing the Met4-TFIID binding surface near the $3'$ edge of the complex. Domain mapping from yeast two-hybrid experiments identified several protein interaction domains on Met4 [4]. The transcriptional activation domain (residues $95-144$) is closer to the Met31 interaction domain (residues $374-403$) than the Cbf1/Met28 interaction domain (residues $616-666$) in one-dimension, but these domains are far apart, so their relative positioning in the native form of Met4 could be different. Based on our findings, we suggest that the relative positioning is the same.

It appears as though either orientation of Met31 can be used within the docking complex (Fig. 2). Stabilization of Met4 has been shown when Met31 has the inverted orientation [4]. For Met31 to be able to stabilize Met4 with either orientation, it must either have two Met4 interaction surfaces, or it has a centrally located interaction surface that is accessible no matter what orientation it binds (*i.e.* flexible). Interestingly, the top 5 genes in our ranking have an inverted Met31 site

(Fig. 6). The total information of these sites might be high because stronger Cbf1 and Met31 sites may be necessary to compensate for the strain of the inverted orientation, but could decrease once we take into account the orientation surprisal.

The maximum distance between the Met4 docking site, as measured to the zero coordinate of the Met31 site, and the gene start is 450 bases. As the TFIID binding site is not at the gene start, this distance is farther than the maximum allowed distance between Met4 and the polymerase. It is difficult to determine the distance of the Met4 docking complex to the transcriptional start since the starts have not been biochemically proven, and computationally it is difficult to predict transcription initiation because of the varied modes of initiation by the polymerase [2]. Basehoar *et al.* found an enrichment of TATAs between 50 and 200 bases upstream of the translational start [32]. This could explain why we did not observe any sites within 100 bases of the gene start.

The spacing range between the Cbf1 and Met31 site is 13 to 68 bases, as determined in Fig. 4. Presumably, it could be larger than this, but our observed range is limited by the spacings in our top 20 hits. DNase I footprint analysis of the Cbf1-Met28 complex showed protection up to position $+8$ on the top strand relative to our Cbf1 zero position in Fig. 1A [16]. If we assume that Met31 can not contact any residues occluded by the Cbf1-Met28 complex due to steric hindrance, then the closest allowed spacing between the Met31 and Cbf1 zero positions would be 13 bases, since the Met31 zero position is the fifth conserved base in our model (Fig. 1B). This is the lower bound determined by our analysis (gene YJL060W in Fig. 6).

A minimum spacing of 13 is also consistent with our observed optimal helical phasing of $-13.24$ bases (Fig. 5). This would place the closest Cbf1 site almost exactly on the same face as its respective Met31 site, one helical turn away. A maximum spacing of 68 bases would correspond to 6 helical turns according to our phasing. The relatively high positioning of these spacings on the helical accessibility curve suggests that docking of Met4 is dependent upon the helical phasing of DNA.

The experimentally determined range by Chiang *et al.* was 21 to 53 bases according to our numbering system [31]. Unfortunately, spacings as large as 68 bases were not tested experimentally. The experimentally determined minimum spacing of 21 is much larger than the minimum we found here. Interestingly, only the "inverted" orientation of Met31 was tested, whereas the shortest distance in this paper corresponds to a Met31 site with the opposite orientation. If helical phasing of the sites is important, then the orientation of Met31 may be more constrained at shorter distances, and this may account for the disparity between the experimentally and computationally determined minimums.

When the constraints inferred from our analysis were imposed on the cooperative binding of Cbf1 and Met31, our ability to predict Met4 regulation was high. Of the top 20 ranked genes in *S. cerevisiae* (according to our model), 18 had the expected microarray expression pattern for Met4 regulation. Many of the sites had also been previously characterized as sulfur utilization genes (Fig. 6). The 2 anomalous genes in the top 20 (Reb1 and Gar1) both had Met31 sites with a "T" instead of "G" at position $+1$ (data not shown), suggesting that this position may be weighted more strongly in a more refined Met31 model. Additionally, nucleosomes could play a large inhibitory

role against spurious combinations of sites, which our model does not account for.

When the microarray data from experiments that affected Met4 binding were directly compared to our information evaluation of each gene (Fig. 7), we saw that almost all genes with Met4 binding sites above 14 bits of information have the expected expression change. This suggests that our approach is giving some reasonable estimate of the energetics of Met4 binding, with a clearly defined threshold for functional binding sites. Presumably, genes that do not have a strong Met4 site, but have the expected microarray data are presumably indirectly regulated. Interestingly the strengths of the Met4 sites are not mainly determined by the strength of Met31 or Cbf1, but by the sum of these sites. This suggests that for cooperatively acting binding sites, a decrease in strength for one site can be compensated for by an increase in strength of the other. Compensation for a decrease in the strength of one binding site by increasing the affinity for another site has been shown experimentally for activation of Pol II by the Epstein-Barr virus protein ZEBRA [33].

Our relative site strength for a given Met4 docking complex is merely the sum of its Cbf1 and Met31 binding sites. We cannot determine the individual spacing effect on binding because we had few sites covering a large spacing range, and our model did well without taking into account varied gap surprisals. If we did have these individual effects, we would expect to see an improved correlation between our information value and the relative expression difference. What we can draw from this analysis is that the sum of individual energetic components, as determined by an information theory approach, gives a reasonably accurate model of a multimeric complex.

It is difficult to say what energetic effect the gap surprisal is representing. It could be related to strain at protein-protein interaction surfaces or strain due to DNA bending. What we are suggesting is that the conserved spacings observed are indicative of the energetics of the system, this of course is the same tenant which seems reasonable in information theory analysis of single transcription factor binding sites [24, 20, 25]. The gap surprisal that we are calculating is most likely a sum of several energetic strains associated with spacing.

We could have determined these physical constraints by clustering co-regulated genes and training the rules of binding for their regulators. The drawback of this approach is that it is not obvious which genes are directly and indirectly regulated, and a given gene may or may not have a binding site. Our approach selects only for genes that are directly regulated, and does not exclude sites that have poor expression data due to experimental error. We are also optimizing our model against all genes in the genome, so we are selecting for a model that represents Met4 binding well, in that it can identify a small subset of sites from all sites in the genome. Presumably the optimal binding site, based on the flexible information theory approach, is the most stable site and the easiest to crystalize. These results could be used to guide crystallographic experiments.

The information content of a given DNA-binding protein ($R_{sequence}$) is a function of the variability within its binding targets [24]. A more stringent binder would have a higher information content, since the variability in its binding targets would be smaller. To be able to distinguish $\gamma$ binding sites within a random DNA of some length **G**, those sites must have an $R_{frequency} = -log_2(\gamma/G)$ bits of information to be identified [24]. It has been shown for many systems that $R_{sequence}$ converges to $R_{frequency}$ [24, 34]. This suggests that if the size of the genome increases and the number

of binding sites remains constant, the information of those sites would have to increase in order to be distinguishable.

As eukaryotic genomes are generally larger than prokaryotic genomes, the amount of information needed to identify $\gamma$ sites would have to be greater. This can be achieved either by increasing the information content of a single factor, or by using multiple factors combinatorially.

Assuming no individual spacing or orientation preferences, the information for this system would be $12.9 + 11.9 - 5.8 - 1 = 18.0$ bits according to equation (4). This would correspond to 1 site every $2^{18.0} = 2.62 \times 10^5$ bases, or about 98 times in the *S. cerevisiae* genome of length 12.8 MB. Our calculation is the number of sites in $2\times$ the genome length, since the complex could associate with either strand. This is a reasonable number of genes according to known sulfur assimilation genes ( $> 20$ genes) [5], the number of predicted regulated genes based on expression difference due to $Cd^{2+}$ treatment (66 genes) [27], and multiple sites per gene as seen in several cases. This suggests that like single acting transcription factors, the information contained in combinatorial binders is related to the $R_{frequency}$ for that system [24]. Others have suggested that this relationship will be maintained for cooperatively acting factors [35]. Interestingly, this information is distributed through individual binding components, as well as the spacing between components, and if one component changed, the others would have to compensate accordingly. This is a complicated process, since Cbf1 can also function independently of Met4 and Met31 [36].

# 6 Acknowledgements

We would like to thank Tom Schneider, Richard Lusk, James Fraser and Virgil Rhodius for comments and useful discussion. This work was supported by the Director, Office of Science, of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231 and NIH grant #RHG002779A.

# References

[1] M. Levine and R. Tjian. Transcription regulation and animal diversity. *Nature*, 424:147–151, 2003.

[2] T. I. Lee and R. A. Young. Transcription of eukaryotic protein-coding genes. *Annu Rev Genet*, 34:77–137, 2000.

[3] T. I. Lee, N. J. Rinaldi, F. Robert, D. T. Odom, Z. Bar-Joseph, G. K. Gerber, N. M. Hannett, C. T. Harbison, C. M. Thompson, I. Simon, J. Zeitlinger, E. G. Jennings, H. L. Murray, D. B. Gordon, B. Ren, J. J. Wyrick, J. B. Tagne, T. L. Volkert, E. Fraenkel, D. K. Gifford, and R. A. Young. Transcriptional regulatory networks in Saccharomyces cerevisiae. *Science*, 298:799–804, 2002.

[4] P. L. Blaiseau and D. Thomas. Multiple transcriptional activation complexes tether the yeast activator Met4 to DNA. *EMBO J*, 17:6327–6336, 1998.

[5] D. Thomas and Y. Surdin-Kerjan. Metabolism of sulfur amino acids in Saccharomyces cerevisiae. *Microbiol Mol Biol Rev*, 61:503–532, 1997.

[6] D. Y. Chiang, A. M. Moses, M. Kellis, E. S. Lander, and M. B. Eisen. Phylogenetically and spatially conserved word pairs associated with gene-expression changes in yeasts. *Genome Biol*, 4:R43, 2003.

[7] R. K. Shultzaberger, L. R. Roberts, I. G. Lyakhov, I. A. Sidorov, A. G. Stephen, R. J. Fisher, and T. D. Schneider. Correlation between binding rate constants and individual information of *E. coli* Fis binding sites. *Nucleic Acids Res., in press*, ■:■–■, 2007.

[8] I. A. Udalova, R. Mott, D. Field, and D. Kwiatkowski. Quantitative prediction of NF-κB DNA-protein interactions. *Proc. Natl. Acad. Sci. USA*, 99:8167–8172, 2002.

[9] R. K. Shultzaberger, R. E. Bucheimer, K. E. Rudd, and T. D. Schneider. Anatomy of *Escherichia coli* Ribosome Binding Sites. *J. Mol. Biol.*, 313:215–228, 2001. http://www.ccrnp.ncifcrf.gov/˜toms/paper/flexrbs/.

[10] R. K. Shultzaberger, Zehua Chen, Karen A. Lewis, and T. D. Schneider. Anatomy of *Escherichia coli* $\sigma^{70}$ promoters. *Nucleic Acids Res.*, 35:771–788, 2007. http://www.ccrnp.ncifcrf.gov/˜toms/paper/flexprom/.

[11] H. Chen, M. Bjerknes, R. Kumar, and E. Jay. Determination of the optimal aligned spacing between the Shine-Dalgarno sequence and the translation initiation codon of *Escherichia coli* mRNAs. *Nucleic Acids Res.*, 22:4953–4957, 1994.

[12] J. Rinke-Appel, N. Junke, R. Brimacombe, I. Lavrik, S. Dokudovskaya, O. Dontsova, and A. Bogdanov. Contacts between 16S ribosomal RNA and mRNA, within the spacer region separating the AUG initiator codon and the Shine-Dalgarno sequence; a site-directed cross-linking study. *Nucleic Acids Res.*, 22:3018–3025, 1994.

[13] D. K. Hawley and W. R. McClure. Compilation and analysis of *Escherichia coli* promoter DNA sequences. *Nucleic Acids Res.*, 11:2237–2255, 1983.

[14] W. R. McClure. Mechanism and control of transcription initiation in prokaryotes. *Annu. Rev. Biochem.*, 54:171–204, 1985.

[15] D. Thomas, I. Jacquemin, and Y. Surdin-Kerjan. MET4, a leucine zipper protein, and centromere-binding factor 1 are both required for transcriptional activation of sulfur metabolism in Saccharomyces cerevisiae. *Mol Cell Biol*, 12:1719–1727, 1992.

[16] L. Kuras, R. Barbey, and D. Thomas. Assembly of a bZIP-bHLH transcription activation complex: formation of the yeast Cbf1-Met4-Met28 complex is regulated through Met28 stimulation of Cbf1 DNA binding. *EMBO J*, 16:2441–2451, 1997.

[17] P. L. Blaiseau, A. D. Isnard, Y. Surdin-Kerjan, and D. Thomas. Met31p and Met32p, two related zinc finger proteins, are involved in transcriptional regulation of yeast sulfur amino acid metabolism. *Mol Cell Biol*, 17:3640–3648, 1997.

[18] G. Wieland, P. Hemmerich, M. Koch, T. Stoyan, J. Hegemann, and S. Diekmann. Determination of the binding constants of the centromere protein Cbf1 to all 16 centromere DNAs of Saccharomyces cerevisiae. *Nucleic Acids Res*, 29:1054–1060, 2001.

[19] T. D. Schneider and R. M. Stephens. Sequence logos: A new way to display consensus sequences. *Nucleic Acids Res.*, 18:6097–6100, 1990. http://www.ccrnp.ncifcrf.gov/˜toms/paper/logopaper/.

[20] T. D. Schneider. Information content of individual genetic sequences. *J. Theor. Biol.*, 189(4):427–441, 1997. http://www.ccrnp.ncifcrf.gov/˜toms/paper/ri/.

[21] T. D. Schneider. Reading of DNA sequence logos: Prediction of major groove binding by information theory. *Meth. Enzym.*, 274:445–455, 1996. http://www.ccrnp.ncifcrf.gov/˜toms/paper/oxyr/.

[22] T. L. Bailey and C. Elkan. The value of prior knowledge in discovering motifs with MEME. *Proc Int Conf Intell Syst Mol Biol*, 3:21–29, 1995.

[23] M. Tribus. *Thermostatics and Thermodynamics*. D. van Nostrand Company, Inc., Princeton, N. J., 1961.

[24] T. D. Schneider, G. D. Stormo, L. Gold, and A. Ehrenfeucht. Information content of binding sites on nucleotide sequences. *J. Mol. Biol.*, 188:415–431, 1986. http://www.ccrnp.ncifcrf.gov/˜toms/paper/schneider1986/.

[25] T. D. Schneider. Theory of molecular machines. II. Energy dissipation from molecular machines. *J. Theor. Biol.*, 148:125–137, 1991. http://www.ccrnp.ncifcrf.gov/˜toms/paper/edmm/.

[26] A. P. Gasch, P. T. Spellman, C. M. Kao, O. Carmel-Harel, M. B. Eisen, G. Storz, D. Botstein, and P. O. Brown. Genomic expression programs in the response of yeast cells to environmental changes. *Mol Biol Cell*, 11:4241–4257, 2000.

[27] M. Fauchon, G. Lagniel, J. C. Aude, L. Lombardia, P. Soularue, C. Petat, G. Marguerie, A. Sentenac, M. Werner, and J. Labarre. Sulfur sparing in the yeast proteome in response to sulfur demand. *Mol Cell*, 9:713–723, 2002.

[28] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A*, 95:14863–14868, 1998.

[29] T. D. Schneider. Strong minor groove base conservation in sequence logos implies DNA distortion or base flipping during replication and transcription initiation. *Nucleic Acids Res.*, 29(23):4881–4891, 2001. http://www.ccrnp.ncifcrf.gov/˜toms/paper/baseflip/.

[30] R. K. Niedenthal, M. Sen-Gupta, A. Wilmen, and J. H. Hegemann. Cpf1 protein induced bending of yeast centromere DNA element I. *Nucleic Acids Res*, 21:4726–4733, 1993.

[31] D. Y. Chiang, D. A. Nix, R. K. Shultzaberger, A. P. Gasch, and M. B. Eisen. Flexible promoter architecture requirements for coactivator recruitment. *BMC Mol Biol*, 7:16, 2006.

[32] A. D. Basehoar, S. J. Zanton, and B. F. Pugh. Identification and distinct regulation of yeast TATA box-containing genes. *Cell*, 116:699–709, 2004.

[33] A. M. Lehman, K. B. Ellwood, B. E. Middleton, and M. Carey. Compensatory energetic relationships between upstream activators and the RNA polymerase II general transcription machinery. *J Biol Chem*, 273:932–939, 1998.

[34] T. D. Schneider. Evolution of biological information. *Nucleic Acids Res.*, 28(14):2794–2799, 2000. http://www.ccrnp.ncifcrf.gov/˜toms/paper/ev/.

[35] D. GuhaThakurta and G. D. Stormo. Identifying target sites for cooperatively binding factors. *Bioinformatics*, 17:608–621, 2001.

[36] N. A. Kent, S. M. Eibert, and J. Mellor. Cbf1p is required for chromatin remodeling at promoter-proximal CACGTG motifs in yeast. *J Biol Chem*, 279:27116–27123, 2004.

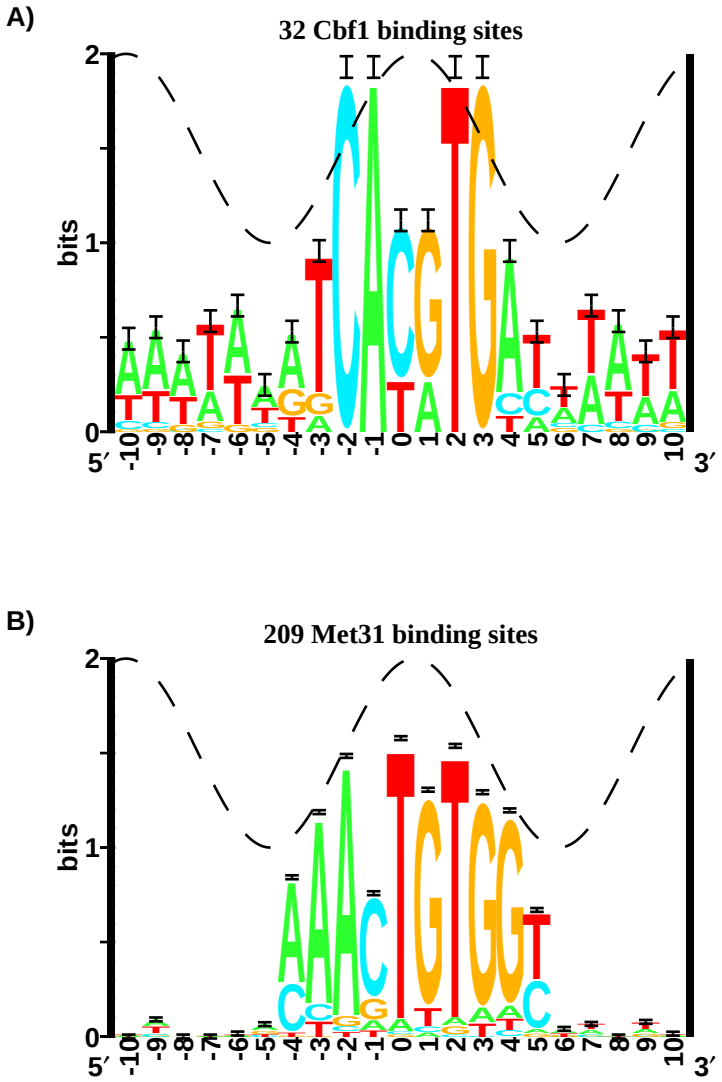



Figure 1: Cbf1 and Met31 sequence logos.

Sequence logos were made as described in Materials and Methods. The height of each letter is proportional to the frequency of that base at that position. The height of the letter stack is the information content at that position. The cosine wave represents the helical twist of B-form DNA. The sequence logos were generated using the standard **Delila** programs [19, 21].

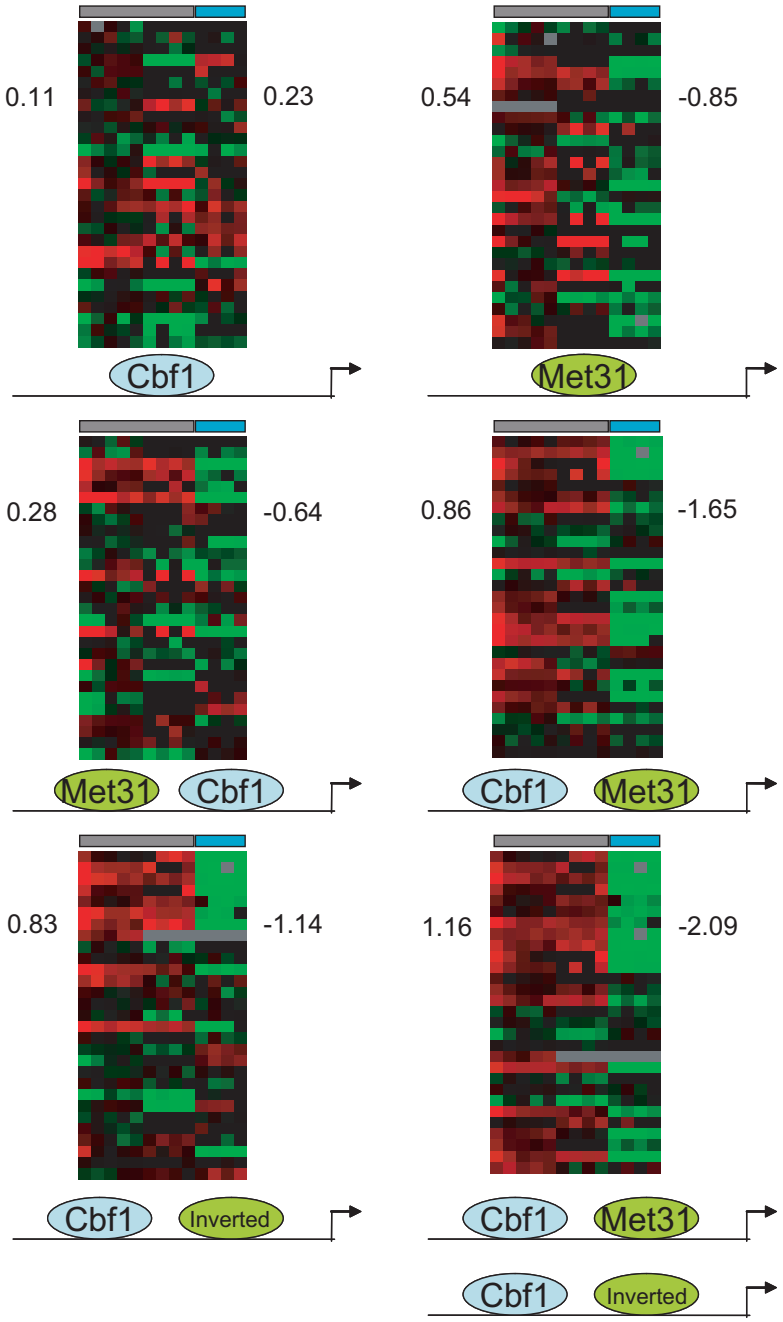



Figure 2: Met4 binding by Cbf1 and Met31 is dependent upon ordering but not orientation. We scanned all intergenic regions in yeast with the models presented in Fig. 1 with different orientations and orderings relative to the gene start point. We then ranked all genes in the genome based on the strength of their strongest upstream binding site, and we present here the corresponding expression changes as determined by microarrays. The experiments that each column represent correspond to those in Fig. 6. For columns $1-9$ (marked with a gray box) we expect regulated genes to have increased expression and therefore to be red. For columns $10-13$ (marked with a blue box) we expect regulated genes to have a decreased expression and therefore to be green. Since the Met31 matrix is asymmetric, it could bind with two different orientations. Those circles labeled "Met31" have the same orientation as the Met31 logo in Fig. 1. Those circles labeled "Inverted" have the opposite orientation (see Fig. 5). The optimal combination in the lower right corner allows for either orientation of Met31. The arrow signifies the gene start. The average expression change for the top 20 genes was calculated for each combination of sites for both the induced (columns 1 to 9) and repressed (column 10 to 13) experiments and are reported next to their respective columns.

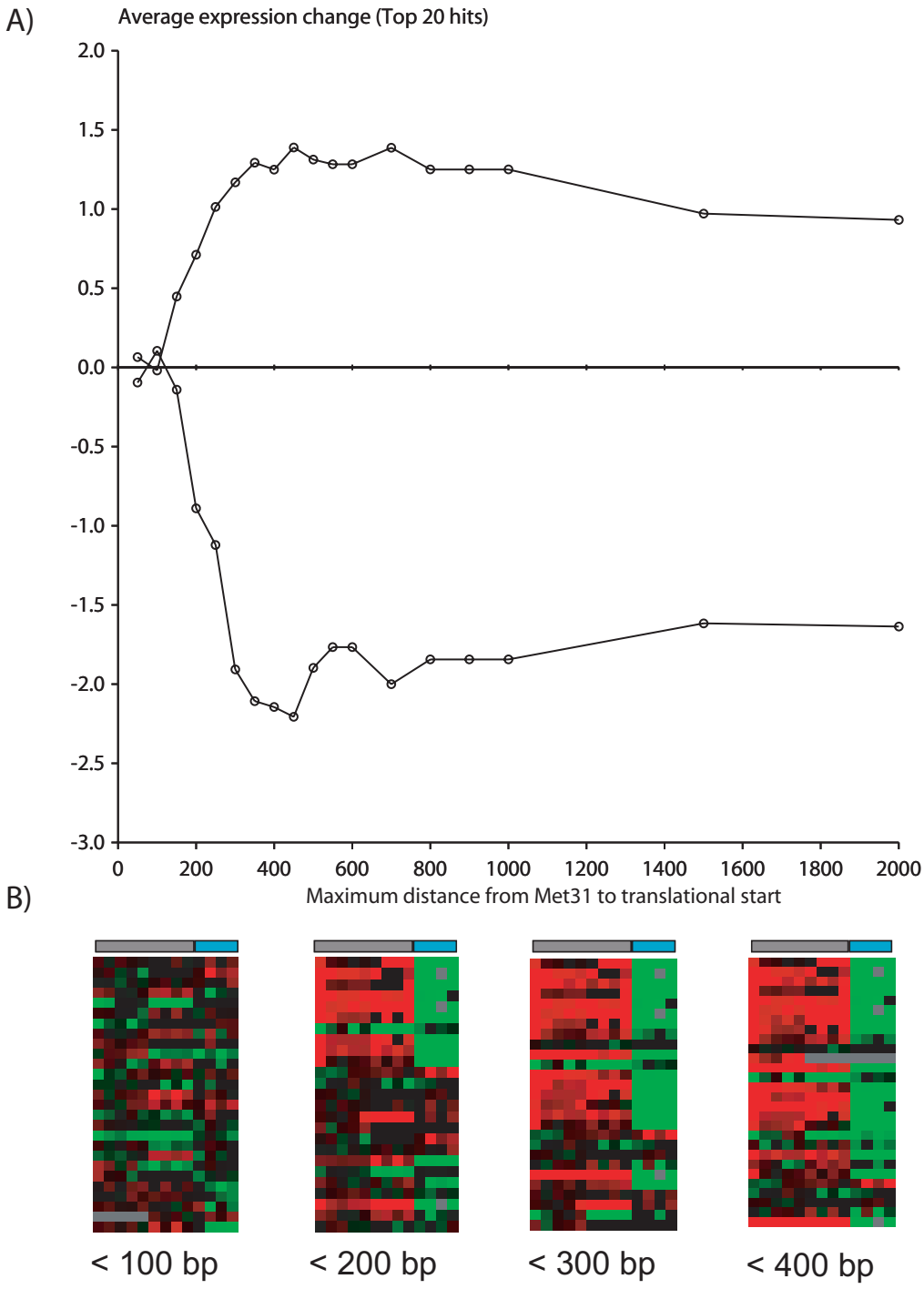



Figure 3: Met4 binding is within 450 bases of the gene start, but not within 100 bases. We varied the allowed distance that the Met31 binding site can be from the gene start point in our models, and quantified how this spacing constraint affected our ability to predict microarray expression data. A) We plotted the average expression change of the top 20 hits in the genome for different maximum spacings from the gene start. The top line corresponds to data from experiments where we expected increased expression (columns 1 to 9 in B), and the lower line is from experiments where we expected decreased expression (columns 10 to 13 in B). The microarray data that corresponds to our gene-ranking are shown in B. The conditions for each column in the microarrays correspond to the labeled columns in Fig. 6.

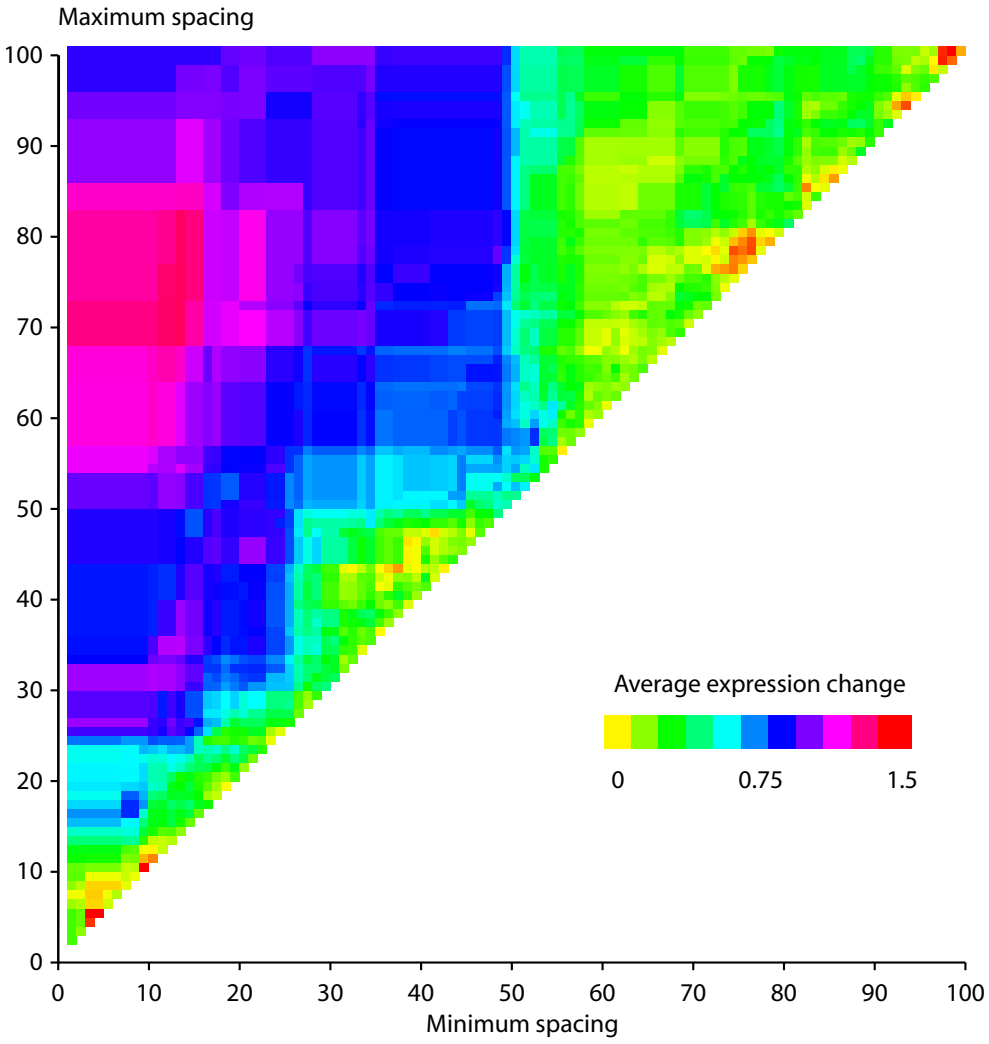



Figure 4: The optimal spacing range between Cbf1 and Met31 is 13 to 68 bases.
We varied the minimum (X axis) and maximum (Y axis) distance that Cbf1 and Met31 could be from each other in our model, and calculated the average expression change within the corresponding top 20 hits, according to these ranges. We show here the average expression change for only those experiments that we expected to have an increased expression (columns 1 to 9 in Fig. 6). The colors correspond to the key in lower portion of the plot.

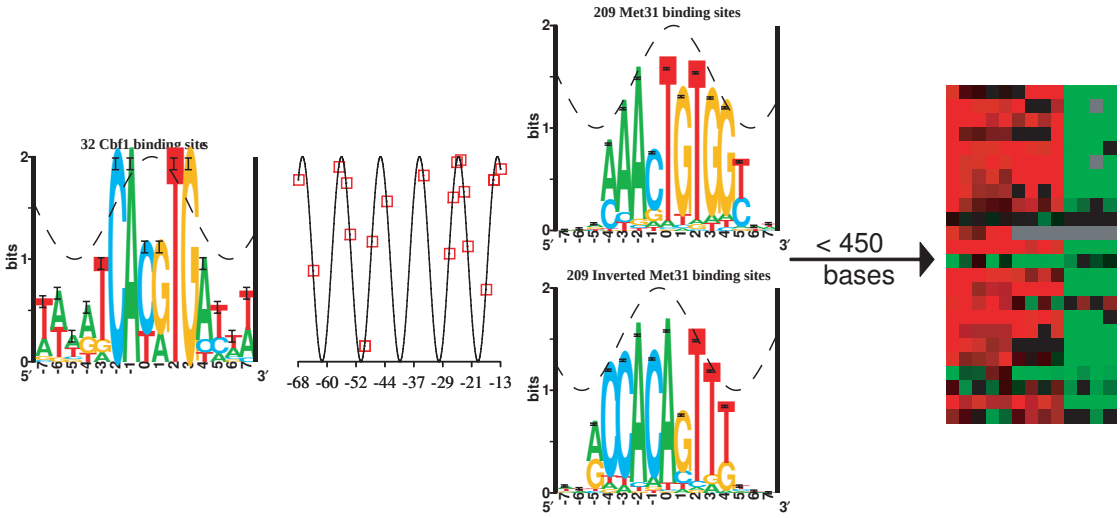
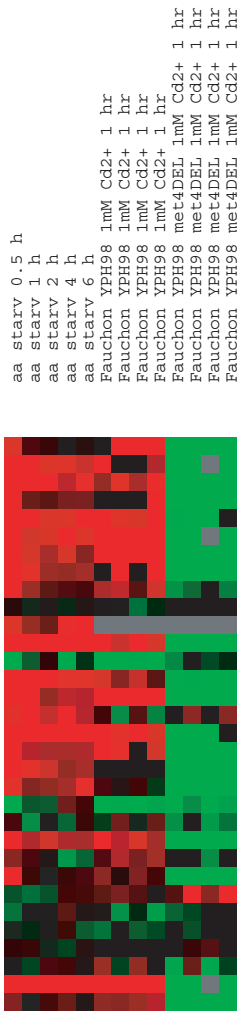



Figure 5: The Met4 activation model based on our analysis.
We summarize here the spacing, ordering, and orientation constraints we used to define functional Met4 binding sites. Since Met31 can bind with either orientation, we show logos for both Met31 orientations. The distances between each set of Cbf1 and Met31 sites were plotted with red boxes on a cosine wave for 20 high-ranking genes to show helical preferences. The arrow represents the translational start, and the allowed distance between the Met4 stabilization complex and the translational start is written above it. The expression data on the right is what was predicted by this model, and is described in Fig. 6.




| Name | Ri | Coord | Or | Cbf1 | ATG | Gene | Function |
|---|---|---|---|---|---|---|---|
| Met30 | 22.99 | 268492 | -1 | -17 | 158 | YIL046W | SULFUR AMINO ACID METBOLISM |
| RAD59 | 21.99 | 345105 | -1 | -44 | 152 | YDL059C | DNA REPAIR AND RECOMBINATION |
| SUL2 | 21.72 | 323262 | -1 | -23 | 283 | YLR092W | TRANSPORT |
| CYS3 | 21.54 | 130615 | -1 | -27 | 187 | YAL012W | METHIONINE BIOSYNTHESIS |
| MET28 | 20.59 | 384265 | -1 | -68 | 149 | YIR017C | SULFUR AMINO ACID METBOLISM |
| MET14 | 20.45 | 439220 | +1 | -34 | 192 | YKL001C | SULFATE ASSIMILATION |
| MET32 | 20.15 | 964883 | -1 | -54 | 322 | YDR253C | METHIONINE METABOLISM |
| MUP3 | 19.55 | 26061 | +1 | -22 | 178 | YHL036W | TRANSPORT |
| YAP5 | 19.08 | 384396 | +1 | -64 | 210 | YIR018W | TRANSCRIPTION |
| REB1 | 17.46 | 337088 | +1 | -19 | 272 | YBR049C | TRANSCRIPTION |
| DDR48 | 17.42 | 608356 | -1 | -55 | 332 | YMR173W | UNKNOWN |
| MET17 | 17.41 | 732271 | +1 | -50 | 273 | YLR303W | METHIONINE BIOSYNTHESIS |
| GAR1 | 17.21 | 283463 | +1 | -27 | 163 | YHR089C | RRNA PROCESSING |
| YIL074C | 16.80 | 222725 | -1 | -57 | 238 | YIL074C | UNKNOWN |
| MET1 | 16.67 | 571003 | +1 | -15 | 251 | YKR069W | METHIONINE BIOSYNTHESIS |
| IDH1 | 16.46 | 559446 | -1 | -15 | 443 | YNL037C | TCA CYCLE |
| MET2 | 16.30 | 117046 | +1 | -48 | 303 | YNL277W | METHIONINE BIOSYNTHESIS |
| YJL060W | 15.91 | 323243 | +1 | -13 | 138 | YJL060W | UNKNOWN |
| MET6 | 15.66 | 342606 | +1 | -50 | 443 | YER091C | METHIONINE BIOSYNTHESIS |
| SAM2 | 15.65 | 1454752 | +1 | -25 | 296 | YDR502C | METHIONINE BIOSYNTHESIS |
| HHF2 | 15.44 | 576370 | +1 | -33 | 358 | YNL030W | CHROMATIN STRUCTURE |
| ARP8 | 15.26 | 592917 | +1 | -16 | 329 | YOR141C | CYTOSKELETON PUTATIVE |
| YHR112C | 15.15 | 335839 | +1 | -26 | 173 | YHR112C | UNKNOWN |
| OPY2 | 14.73 | 697205 | -1 | -22 | 389 | YPR075C | SIGNALING PUTATIVE |
| MXR1 | 14.70 | 234775 | -1 | -68 | 161 | YER042W | OXIDATIVE STRESS RESPONSE |
| ACP1 | 14.67 | 80710 | -1 | -15 | 168 | YKL192C | FATTY ACID BIOSYNTHESIS |
| LCB3 | 14.65 | 158052 | -1 | -50 | 133 | YJL134W | SPHINGOLIPID METABOLISM |
| SFH1 | 14.63 | 778313 | +1 | -47 | 449 | YLR321C | CHROMATIN STRUCTURE |
| PAN1 | 14.53 | 370029 | -1 | -40 | 124 | YIR006C | CYTOSKELETON AND ENDOCYTOSIS |
| SNC2 | 14.20 | 931319 | +1 | -51 | 241 | YOR327C | SECRETION |
| MET10 | 14.09 | 213095 | +1 | -24 | 205 | YFR030W | SULFATE ASSIMILATION |
| YML018C | 14.08 | 236206 | +1 | -25 | 254 | YML018C | UNKNOWN |

Figure 6: The top hits are involved in sulfur amino acid biosynthesis.
These are the top hits according to our optimal spacing values. The first 9 columns are data for experiments that should induce the expression of Met4 regulated genes and give a red pattern. The last 4 columns we expect to see a decrease in expression of Met4 regulated genes and give a green pattern. Experiment information for each column is reported vertically above each column. Each row corresponds to a different gene followed by its common name, its flexible information ($R_i$), the coordinate of the Met31 binding site in the *S. cerevisiae* genome, the orientation of the Met31 matrix, the distance Cbf1 is upstream of Met31, the distance the gene start is downstream of Met31, the gene name according to its annotation in the yeast genome, and a description of its function.

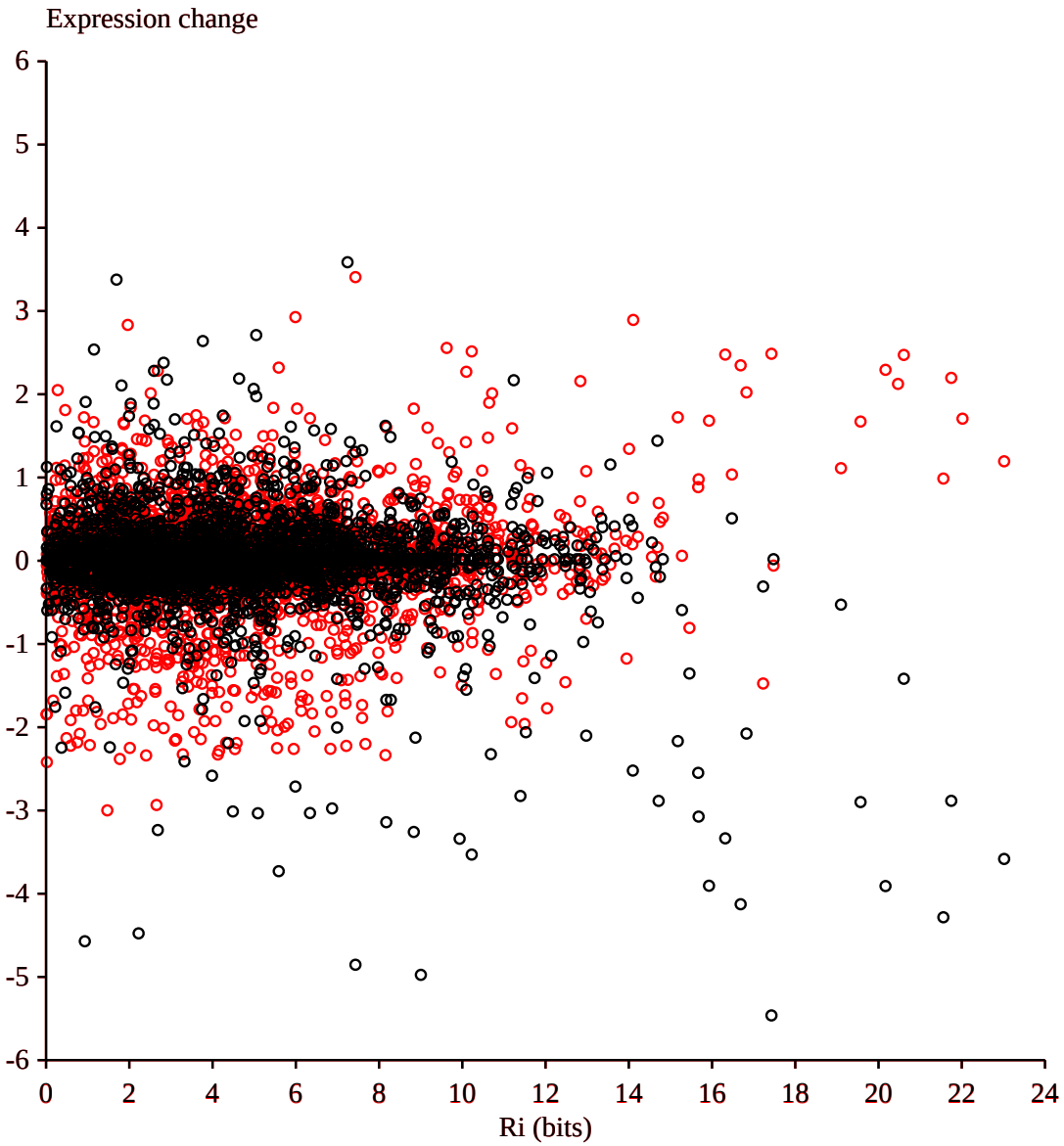



Figure 7: Ri vs. expression change.
The flexible information ($R_i$) of the strongest site for each gene is on the abscissa and the average induction or repression expression change is on the ordinate. For each gene, induction data were averaged from the first nine experiments in Fig. 6 (red circles), and repression data were averaged from the last four experiments in Fig. 6 (black circles).